华东理工大学2019–2020 学年第 二 学期

《多元统计分析》实验报告

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 实验名称 | 实验1 数据整理与描述统计分析 | | | | | | | | | | |
| 专 业 | | 数学与应用数学 | 姓名 | 梁天一 | 学号 | 10172911 | | 组名/组号 | | SCP-173 |
| 实验时间 | | 公元2020年 | | 实验地点 | 电脑前 | | 指导教师 | |  | |

|  |
| --- |
| 实验目的/要求 |
| 1、掌握数据整理的基本方法：观察量排序（Sort Cases）、变量排序(Rank Cases)、计算新的变量(Compute Variables)、拆分数据文件(Split Files) 、分类汇总(Aggregate)等。  2、熟练应用SPSS输出描述统计量和绘制统计图。 |
| 实验内容 |
| 1、对“employee data.sav ”进行数据整理，并分别给出三种工作类别（jobcat）的薪水（salary）的描述统计量（均值、方差等）。  2、对第1章的习题4进行描述统计分析。 |
| 实验总结 |
|  |
| 教师批阅： 实验成绩：  教师签名: 日期： |

学号+姓名+实验一

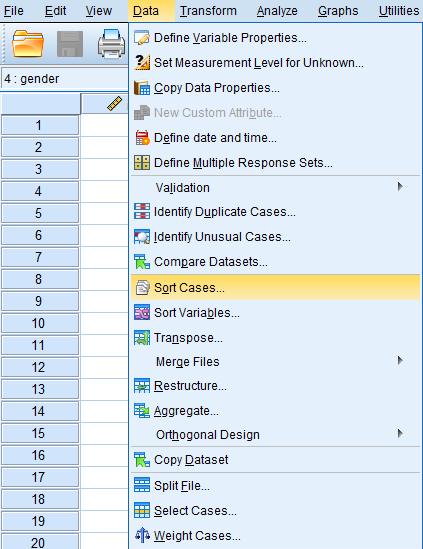
实验报告正文：

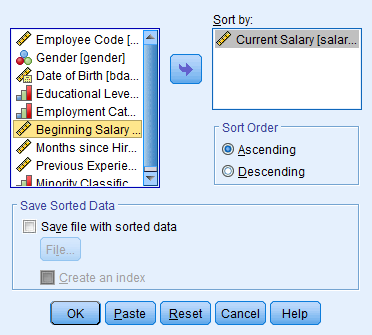
（包括实验步骤、实验结果、实验过程中出现的问题及解决方法等。要求做到内容精炼、层次清楚、格式规范。）

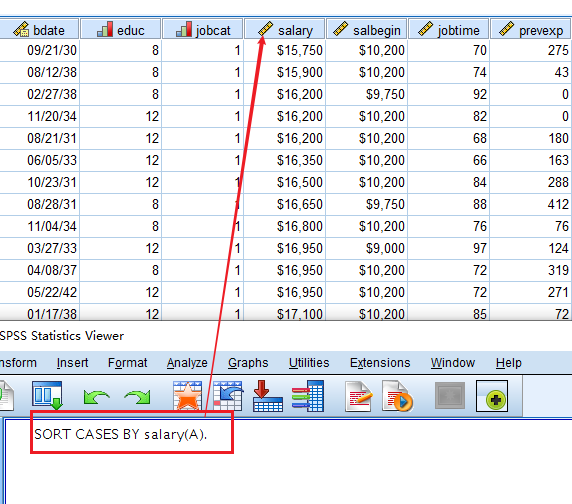
# 对“employee data.sav ”进行数据整理，并分别给出三种工作类别（jobcat）的薪水（salary）的描述统计量（均值、方差等）。

首先我们使用英文版的原汁原味的SPSS EDIT-OPTION-LANGUAGE选择英文, 与教程接轨。

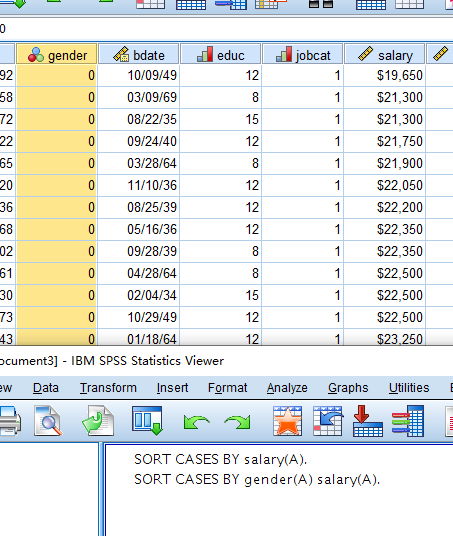
## 观察量排序





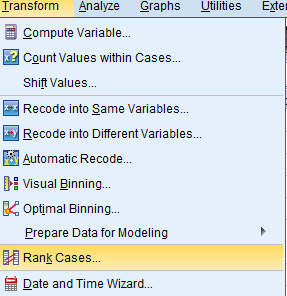


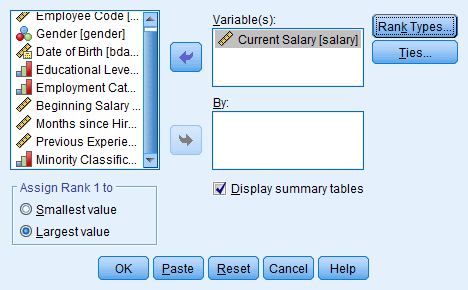
可以看出所有gui命令对有相应的脚本语言,为撰写脚本来批处理提供了方便.这个命令默认参数是让所有的元组根据选取的属性的值从小到大排列.

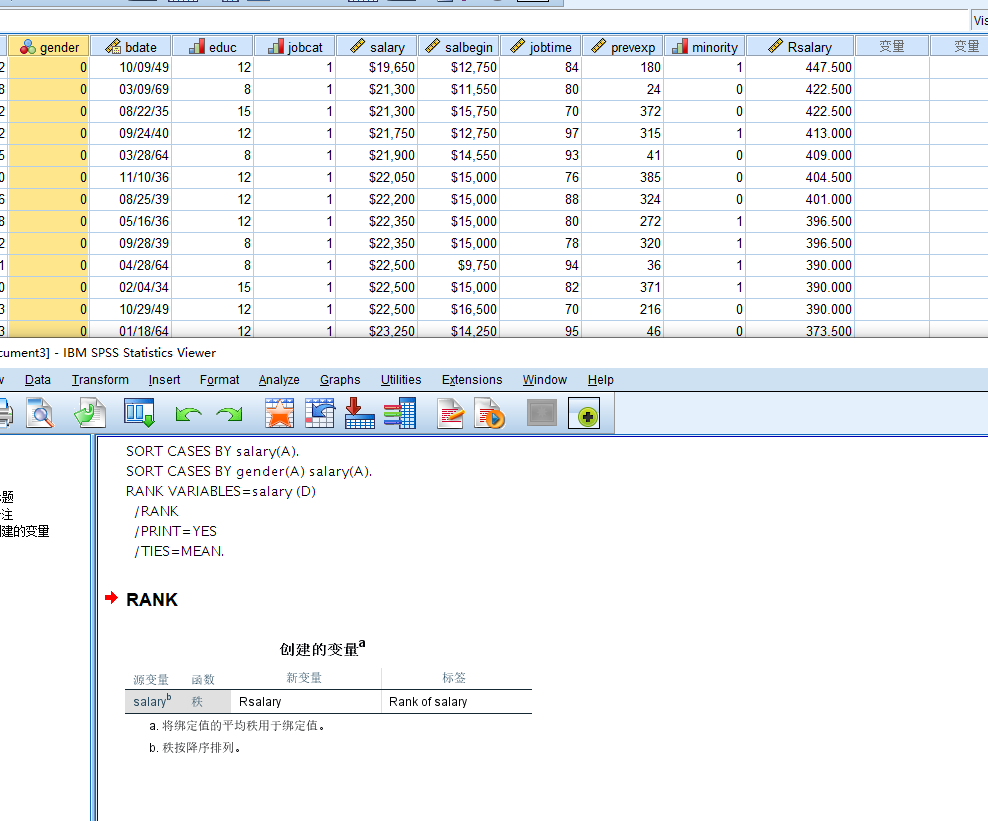


如果是多变量会类似excel中的排序操作 优先级从前往后依次递减。每次先对优先级高的进行排序。

## 变量排序

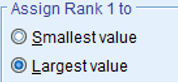


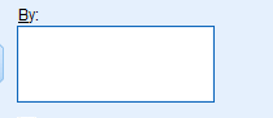




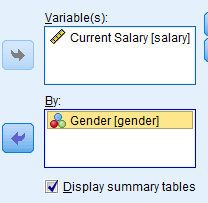
可以看出增加了一个变量用来表示排名.

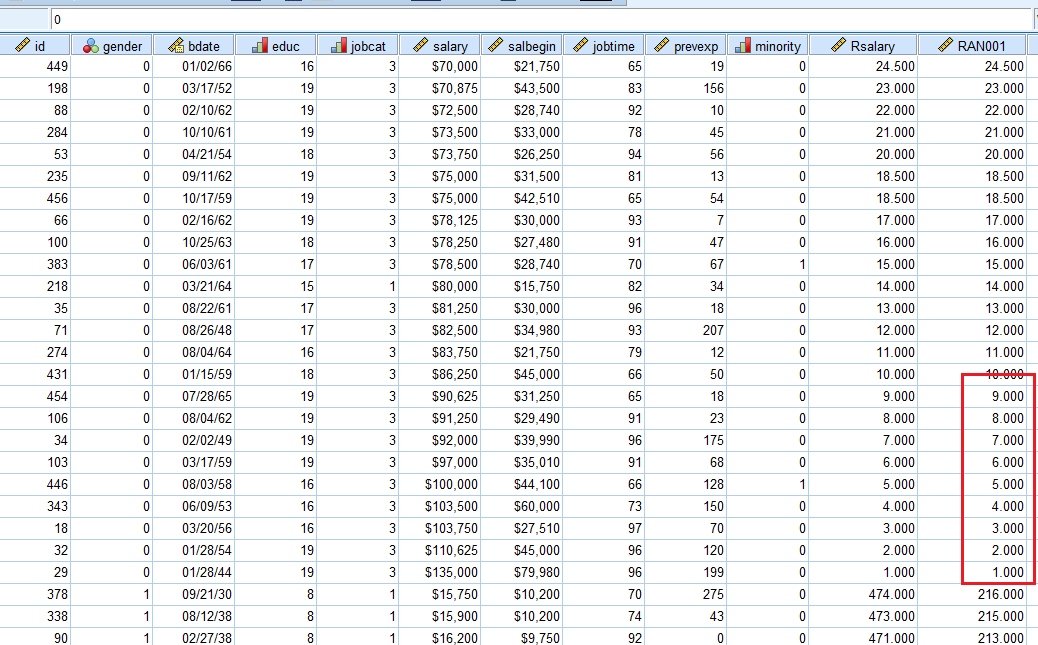
其他参数说明:

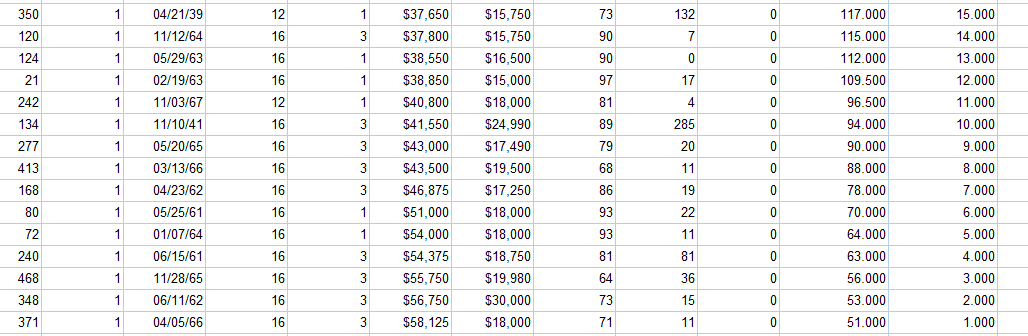
表示排第一的是最大的数,然后依次递减



表示分组排名







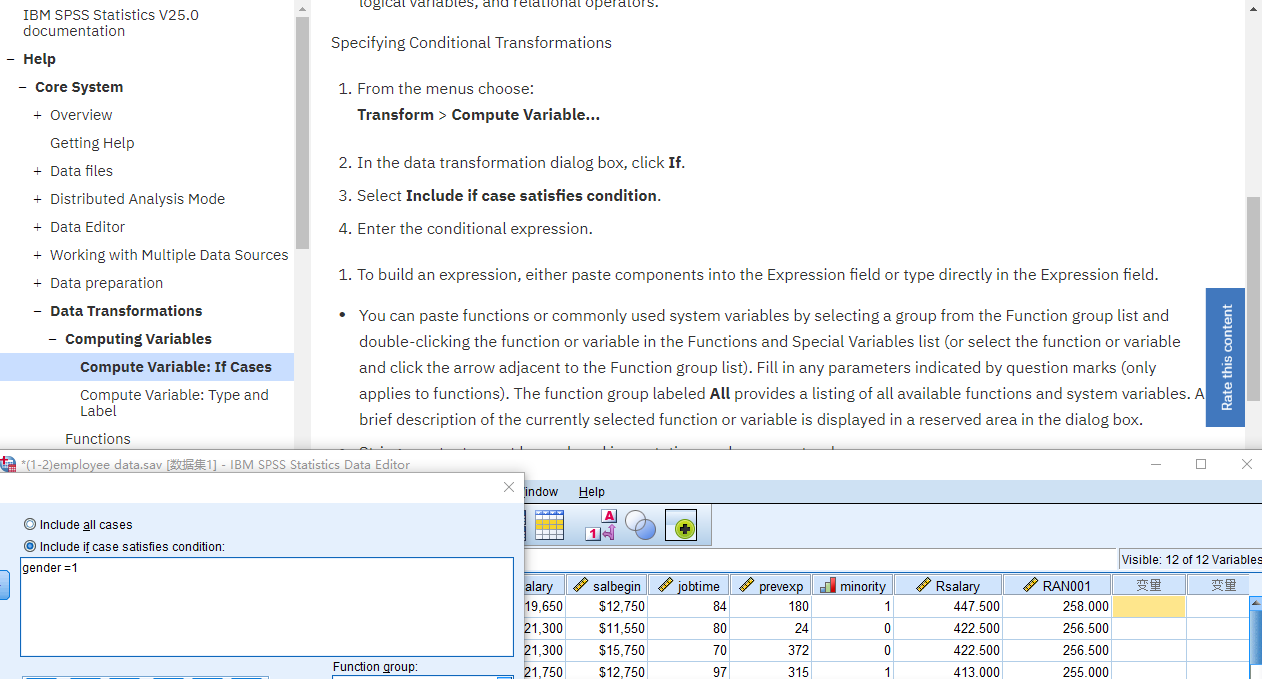
会按照不同分组排名,走出了代替excel进行日常数据处理的第一步

## 计算新的变量

类似于excel的单元格公式

我们用如果gender=0就加薪水100 , gender=1就加薪水1000演示

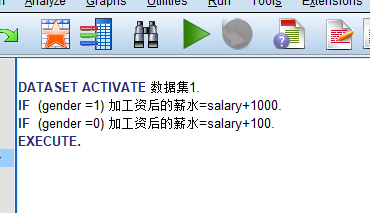
要实现条件语句,这里我们请出知名工具

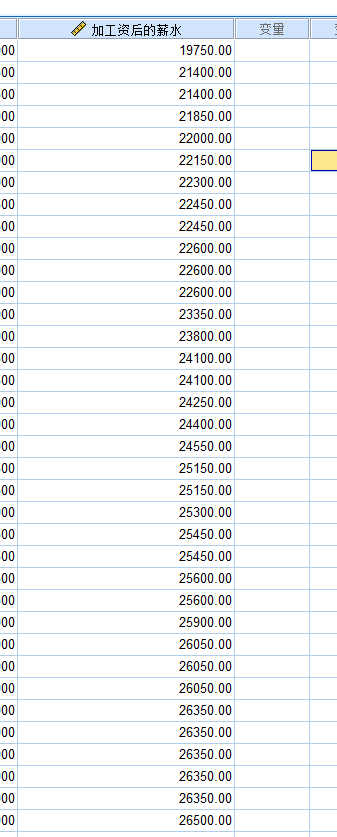
得到

果然啥都看不懂,所以我们进行百度,找到[相关经验](<https://jingyan.baidu.com/article/fcb5aff7ee4068edab4a7110.html>)

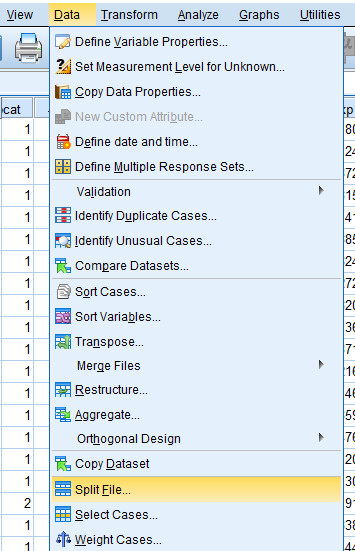
Gui输入脚本之后,我们可以点击paste粘贴代码

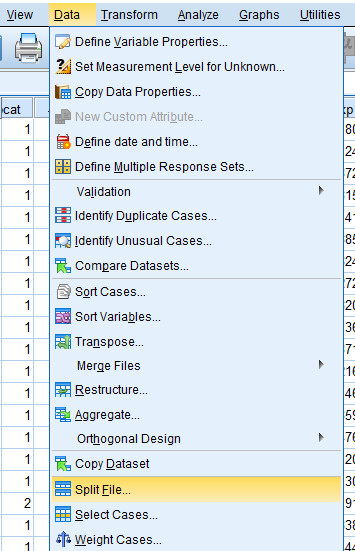
然后写代码

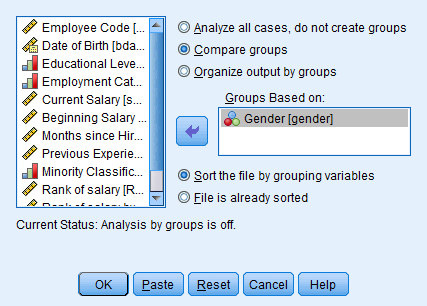


执行成功,这个故事告诉我们,与其点点点,不如写代码,逻辑和windows上点一点gui和shell编程来运维是一样的

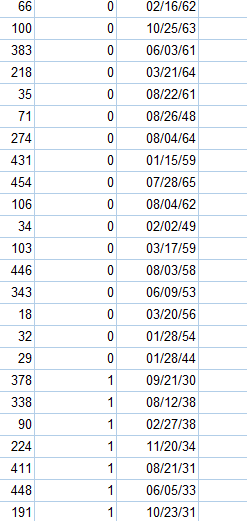
## 拆分数据文件



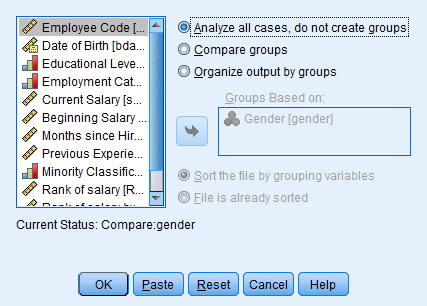




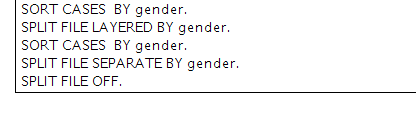
可以看出已经分组了



如果要取消分组,需要

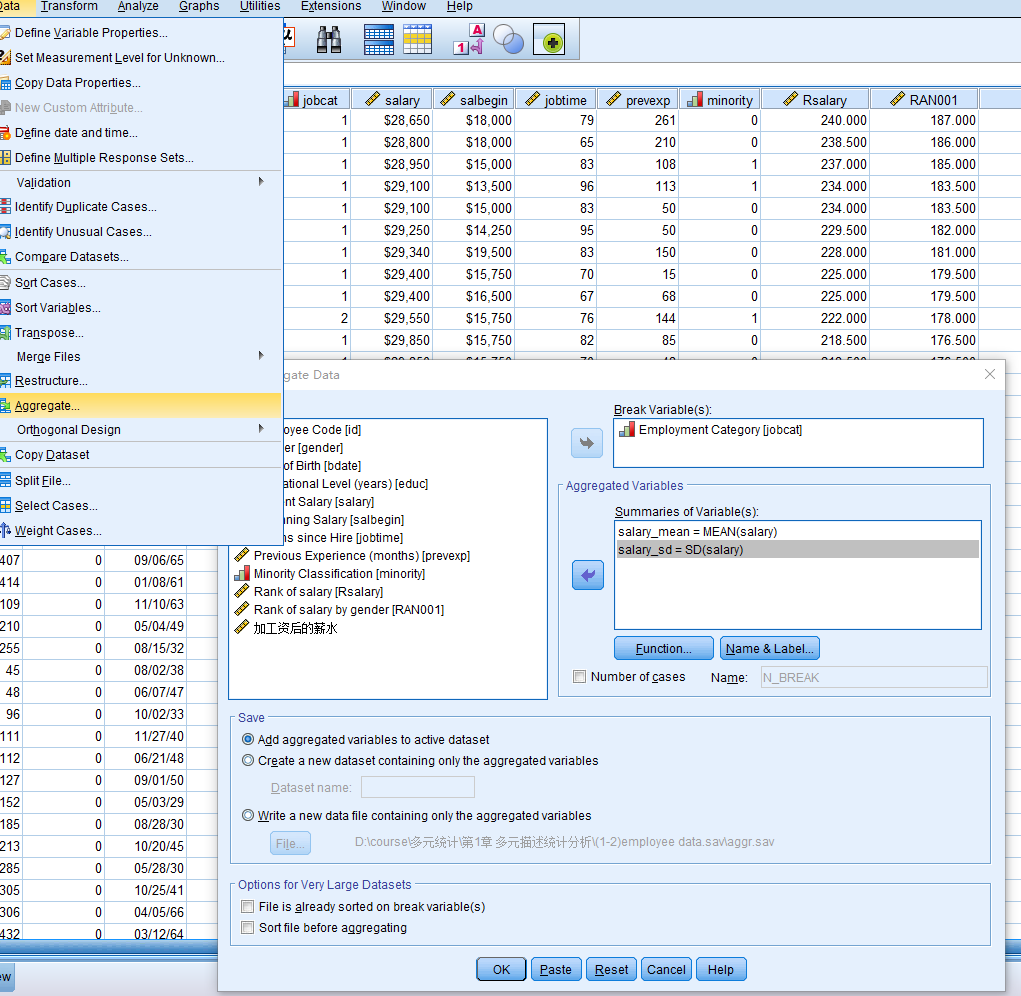


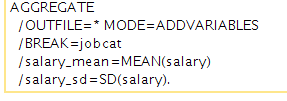
代码可以看出先进行了一次排序

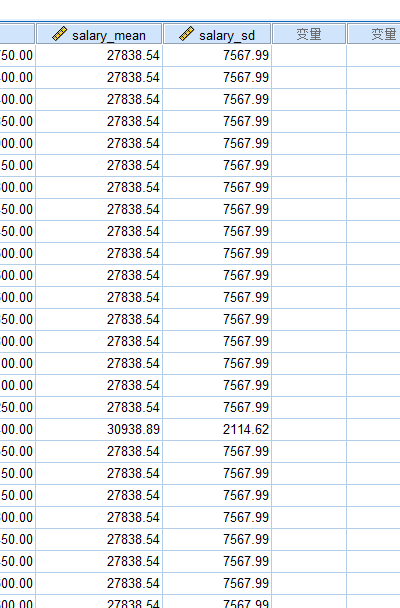


## 分类汇总

这个操作可以直接把下面一步操作做了



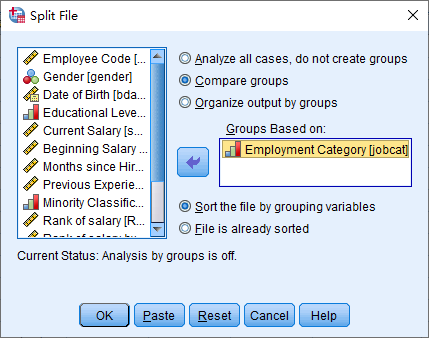


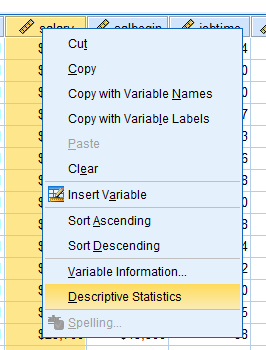
可以看出创建了新的类,根据分组依据算出新的值

## 三种工作类别（jobcat）的薪水（salary）的描述统计量（均值、方差等）。

这个问题已经被分类汇总解决了,但是变成一列比较难看,因此我们看一下另一个操作

分组然后统计描述





得到

SORT CASES BY jobcat.

SPLIT FILE LAYERED BY jobcat.

FREQUENCIES VARIABLES=salary

/STATISTICS=RANGE MINIMUM MAXIMUM STDDEV MEAN MEDIAN

/FORMAT=NOTABLE

/ORDER=ANALYSIS.

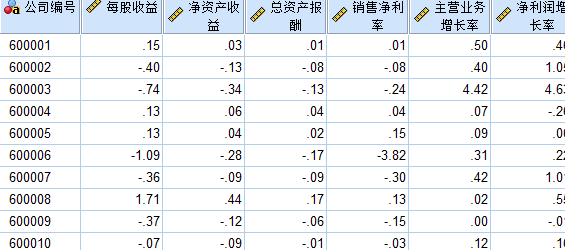
所以分类汇总本质上是设置一个全局后面,来让后面的STATISTICS之前都会SPLIT FILE LAYERED BY jobcat.

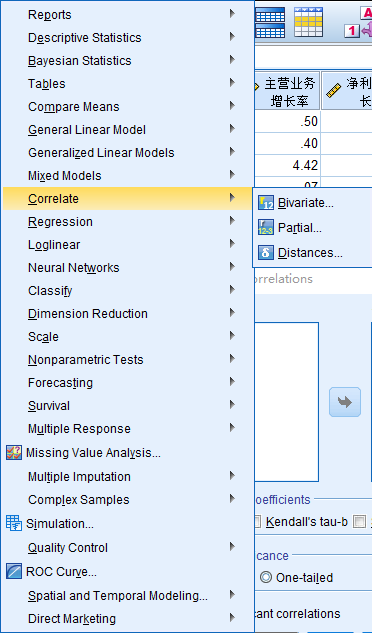
# 2、对第1章的习题4进行描述统计分析。

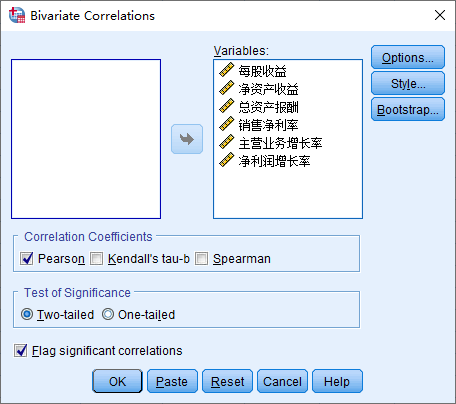
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 每股收益 | 净资产收益率 | 总资产报酬率 | 销售净利率 | 主营业务增长率 | 净利润增长率 |
| 600001 | 0.148 | 0.028 | 0.011 | 0.011 | 0.499 | 0.463 |
| 600002 | -0.399 | -0.127 | -0.078 | -0.079 | 0.404 | 1.046 |
| 600003 | -0.744 | -0.335 | -0.129 | -0.241 | 4.415 | 4.628 |
| 600004 | 0.132 | 0.061 | 0.037 | 0.043 | 0.069 | -0.260 |
| 600005 | 0.131 | 0.041 | 0.023 | 0.148 | 0.087 | 0.001 |
| 600006 | -1.087 | -0.281 | -0.167 | -3.822 | 0.306 | 0.215 |
| 600007 | -0.361 | -0.088 | -0.085 | -0.296 | 0.419 | 1.006 |
| 600008 | 1.712 | 0.435 | 0.167 | 0.132 | 0.024 | 0.553 |
| 600009 | -0.374 | -0.121 | -0.063 | -0.149 | -0.004 | -0.012 |
| 600010 | -0.070 | -0.091 | -0.010 | -0.031 | 0.115 | 0.157 |

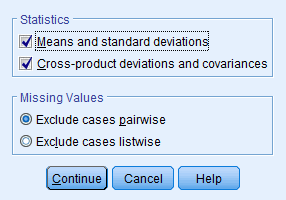
## 试对上述数据计算样本均值矩阵、协方差阵、相关系数等描述统计量。

首先输入数据



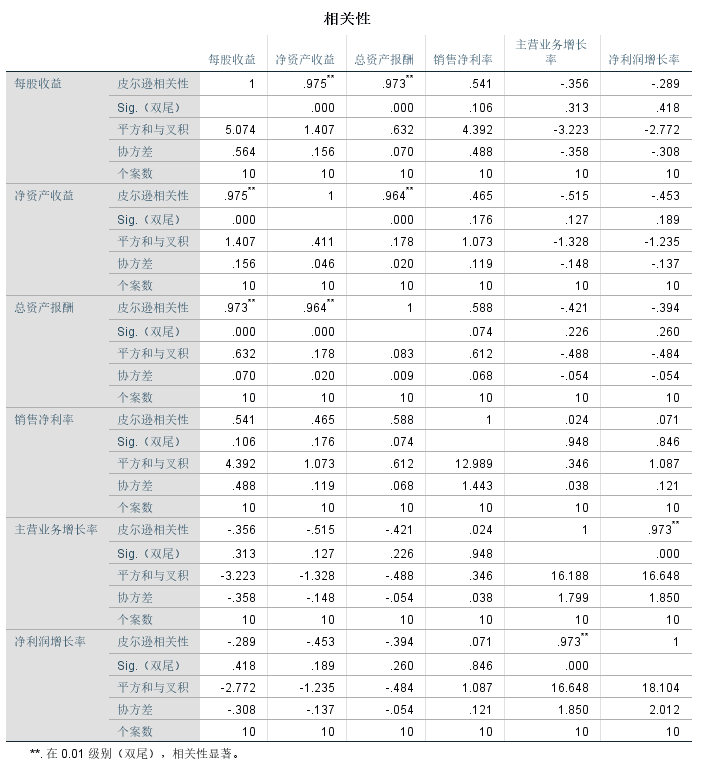






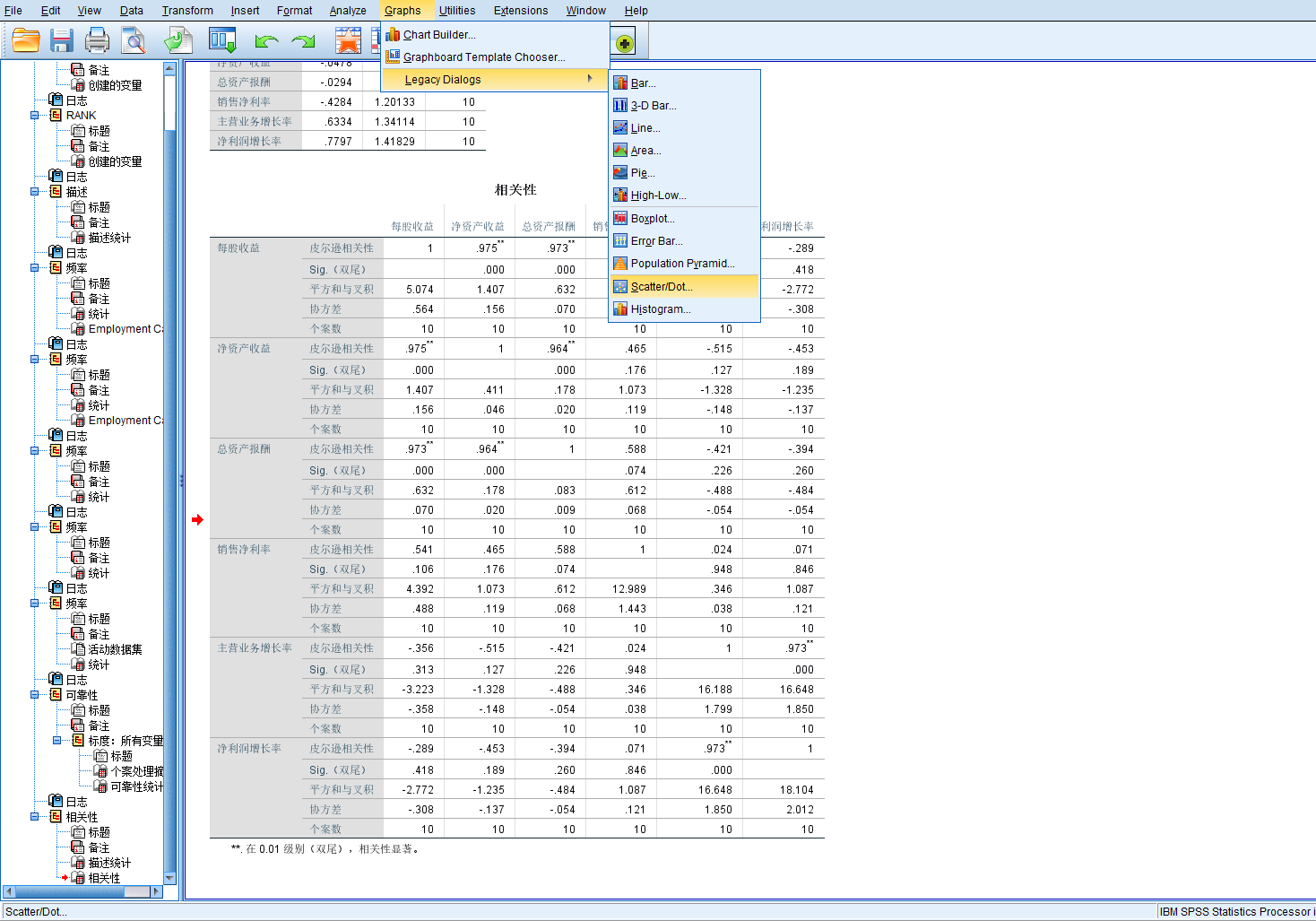
这样就得到了6个属性的样本均值矩阵 协方差矩阵 相关系数

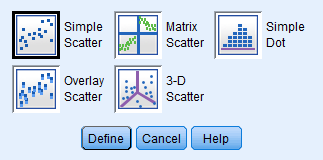




特别注意,这里给出的样本协方差是,此外会自动计算显著性水平0.01的时候相关性显著的结果

## 试描绘上述多个变量的矩阵散点图。



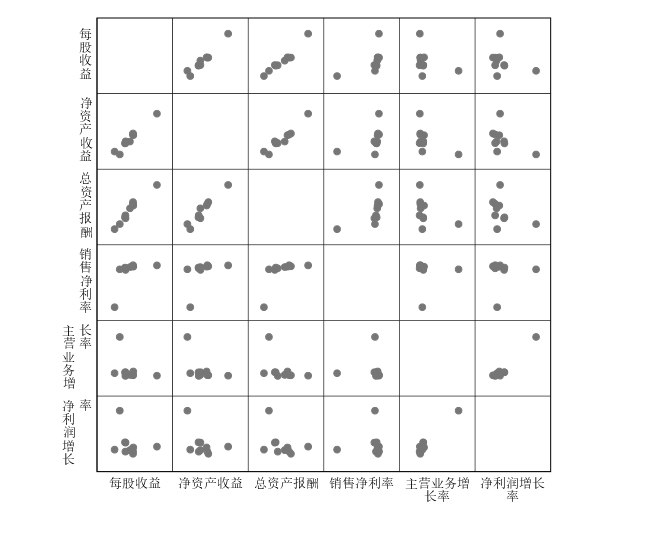


选择第二个,然后把所有变量放进去

GRAPH

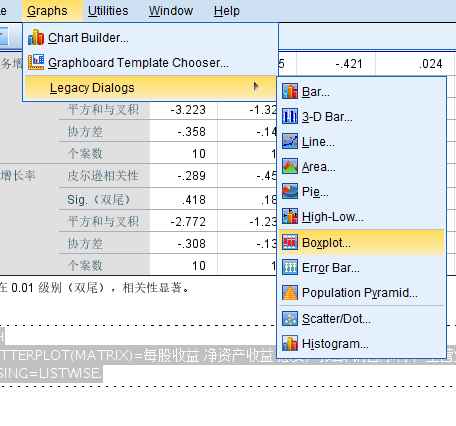
/SCATTERPLOT(MATRIX)=每股收益 净资产收益 总资产报酬 销售净利率 主营业务增长率 净利润增长率

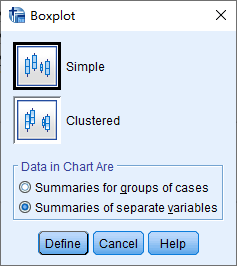
/MISSING=LISTWISE.

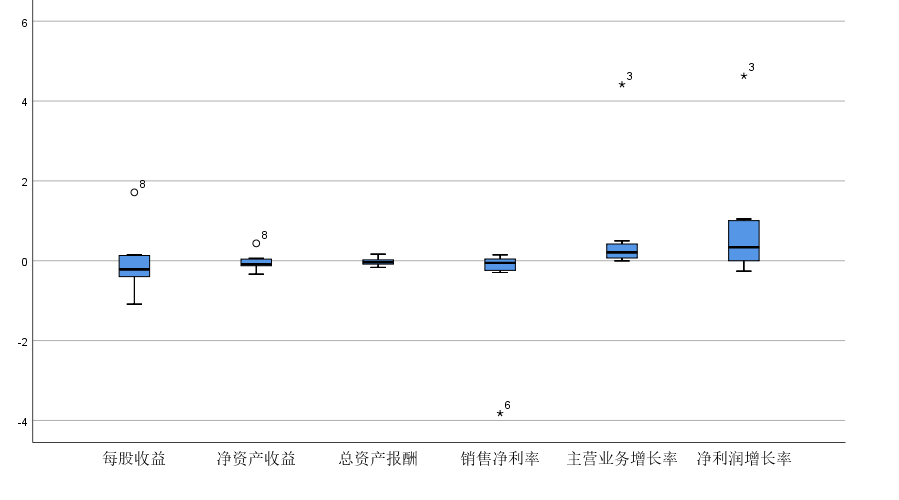
这里

可以发现皮尔逊相关系数越高,就越发像斜线

## 试用多维箱线图描述上述数据



 之后放入所有变量



刻印看出数据大致分布,均值,方差,离群点等

EXAMINE VARIABLES=每股收益 净资产收益 总资产报酬 销售净利率 主营业务增长率 净利润增长率

/COMPARE VARIABLE

/PLOT=BOXPLOT

/STATISTICS=NONE

/NOTOTAL

/MISSING=LISTWISE.

**总结**

所以什么时候做通过从多属性回归分析宝可梦的cp值的实验

了解到gui操作界面与spss脚本语言之间的对应关系, 为通过脚本来批处理提供了可能.

* 分类汇总本质上是设置一个全局后面,来让后面的STATISTICS之前都会SPLIT FILE LAYERED BY jobcat.
* 几乎每个操作都会默认排序一下,所以为了性能效率在大样本下预先sort然后批处理操作

在实验的过程中，我探究了相关工具每个选项的含义。

看见不会就点击help查看官方文档体现了研究人员基本的素养.然后果然官方文档不配图语义不清,不放代码导致没有可执行性,不如谷歌一下.

下一步工作是研究相关代码的修改接口,来适应相关统计方法的更新,或者尝试以下操作